Εργαστηριακή εργασία εξαμήνου

Μηχανική μάθηση

Διδάσκων: Κώστας Γιαννάκης

Προθεσμία υποβολής: Παρασκευή, 16 Ιανουαρίου 2026

Παραδοτέα και βαθμολόγηση

Η εργασία είναι προαιρετική. Το τελικό παραδοτέο θα πρέπει να είναι ένα αρχείο pdf με την αναφορά σας (με τις απαντήσεις σας στις σχετικές ερωτήσεις, την όποια χρήση LLMs κάνατε, screenshots όπου ζητούνται και τα αντίστοιχα plots). Η εργασία είναι προαιρετική και προσφέρει μέχρι 2 επιπλέον βαθμούς στην τελική βαθμολόγηση που μπορεί να φτάσει μέχρι 3 μονάδες με ερωτήματα μεγάλης δυσκολίας για όσες και όσους παρακολουθούσαν τα εργαστήρια.

Το τελικό παραδοτέο θα πρέπει να είναι ένα αρχείο pdf με την αναφορά σας ΚΑΙ ο πηγαίος κώδικας σε ένα ενιαίο script (σε .zip ή .txt). Ο κώδικας θα πρέπει να είναι καλά τεκμηριωμένος με σχόλια (ακολουθώντας τον τρόπο σχολιασμού στον κώδικα που χρησιμοποιούσαμε στα εργαστήρια). Θα πρέπει κατ' ελάχιστο να περιλαμβάνει τα βήματα προεπεξεργασίας (συμπεριλαμβανομένου του βήματος φιλτραρίσματος με βάση τον αριθμό μητρώου σας, δείτε παρακάτω), ενδεικτικά γραφήματα με τις κατανομές για επιλεγμένα χαρακτηριστικά (με χρήση barplot), καθώς και ένα grid plot (hint: το είδαμε σε προηγούμενα εργαστήρια με την εντολή pairs. Εάν έχετε πολλά δεδομένα, επιλέξτε να απεικονίσετε ένα δείγμα παίρνοντας ας πούμε 1000 τυχαίες γραμμές). ΠΡΟΣΟΧΗ: τα όποια αποτελέσματα και γραφήματα καταγράψετε στην αναφορά σας θα πρέπει να μπορούν να αναπαραχθούν αυτούσια από μένα τρέχοντας μονομιάς τον κώδικά σας!!

Σκοπός της εργασίας

Η άσκηση των φοιτητών στους βασικούς αλγόριθμους ταξινόμησης και ομαδοποίησης, καθώς και σε μέτρα απόδοσης.

Σημαντικό!

- Η χρήση LLMs για αντιγραφή της εργασίας απαγορεύεται. Επιτρέπεται η καλόπιστη χρήση τους με άσκηση ερωτήσεων και ξεκάθαρη αναφορά τόσο του μοντέλου LLM που χρησιμοποιήσατε όσο και του κάθε prompt που χρησιμοποιήσατε.
- Η εργασία είναι ατομική και δεν επιτρέπεται η αντιγραφή.
- Μπορείτε και συνιστάται να ανατρέξετε στα προηγούμενα εργαστήρια και τον κώδικά τους.

1ο μέρος (90% του βαθμού της εργασίας)

Οδηγίες εκτέλεσης της εργασίας:

- Κατεβάστε και αποσυμπιέστε το αρχείο ML_tasks.zip. Περιλαμβάνει:
  - το αρχείο mushrooms_clean.csv
  - το αρχείο code.R . Αυτό θα είναι το αρχείο στο οποίο θα πρέπει να δουλέψετε.
- Στο πάνω μέρος του κώδικα, υπάρχει ένα τμήμα στο οποίο διαβάζουμε το συνοδευτικό αρχείο (προσοχή να είναι στο ίδιο φάκελο) και γράφουμε στη μεταβλητή YYY τον αριθμό του μητρώου μας, με προσοχή να μην περιλαμβάνει ελληνικούς χαρακτήρες. Πχ, γράφουμε ‘p2007041’ ή ‘inf2022555’. Δεν αλλάζουμε τίποτα άλλο στο συγκεκριμένο τμήμα.
  - Αυτό οδηγεί στον μοναδικό κατακερματισμό/δειγματοποίηση του αρχικού dataset (YYY.csv) και καταλήγει κάθε φοιτητής με 15000 τυχαία δείγματα από το αρχικό σύνολο που περιελάμβανε >60000 γραμμές. **Αυτό το δείγμα (δηλαδή το dataframe με όνομα sampled_mushrooms και ~15000) είναι μοναδικό για κάθε φοιτητή. Αυτό είναι το τελικό dataset που θα χρησιμοποιήσει ο καθένας και η καθεμία σας.**
- Τα δεδομένα απαιτούν να:
  - Αφού πάρετε το ατομικό σας υποσύνολο με την παραπάνω διαδικασία, χρησιμοποιήστε τις συναρτήσεις str(), summary() και head() για να ελέγξετε τη δομή και τη στατιστική σύνοψη. Καταγράψτε και περιγράψτε με συντομία τα ευρήματά σας.
  - Αναγνωρίστε ποιες μεταβλητές είναι κατηγορικές και ποιες είναι αριθμητικές.

- Απεικονίσετε barplots για κάθε μεταβλητή. Για βοήθεια, ακολουθεί ενδεικτικός κώδικας:

```
# frequency table for Feature_X
featureX_counts <- table(sampled_mushrooms$Feature_X)

# barplot for Feature_X
barplot(featureX_counts, main = "Bar Plot of Feature_X", xlab =
"Feature_X categories", ylab = "Count", col = "skyblue")
```

- Ομοίως, οπτικοποιήστε την κατανομή της μεταβλητής στόχου (στην περίπτωσή μας, η μεταβλητή με όνομα class).
- Απεικονίστε ένα συνολικό grid plot για όλα τα χαρακτηριστικά. Λόγω μεγέθους, χρησιμοποίστε ένα τυχαίο δείγμα 1000 γραμμών από το συνολικό των 15000.
- Σιγουρευτείτε ότι το σύνολο δεδομένων σας δεν έχει διπλότυπα (duplicates), δηλαδή δείγματα με ακριβώς ίδια χαρακτηριστικά.
- Διαχειριστείτε τις ελλείπουσες τιμές (missing values) αν υπάρχουν.
- **Εντοπίσατε ακραίες τιμές; Αν ναι, πώς το διαχειριστήκατε (πχ τις απομακρύνατε? Τις αντικαταστήσατε με μέση ή διάμεσο τιμή);**
- Σκεφτείτε να χρησιμοποιήσετε na.omit() ή τεχνικές αποκατάστασης.
- Μετατρέψτε τις κατηγορικές μεταβλητές σε παράγοντες όπου χρειάζεται χρησιμοποιώντας το as.factor().
- Κανονικοποιήστε τα αριθμητικά χαρακτηριστικά χρησιμοποιώντας το scale() ΑΝ ΕΙΝΑΙ ΑΠΑΡΑΙΤΗΤΟ (με δικαιολόγηση), ειδικά πριν εφαρμόσετε συγκεκριμένες μεθόδους (σκεφτείτε ποιες). Εξηγήστε εάν και πότε χρειάζεται η κανονικοποίηση και δηλώστε, με αιτιολόγηση, εάν εσείς προβήκατε ή όχι σε κανονικοποίηση στο σύνολο δεδομένων σας.
- Για τους αλγόριθμους ταξινόμησης, η κλάση ταξινόμησης (ή αλλιώς η μεταβλητή στόχου ή εξαρτημένη μεταβλητή ή μεταβλητή Υ) είναι αυτή με το όνομα ‘class’
- Για τους αλγόριθμους ομαδοποίησης, επιλέγετε και δουλεύτε αποκλειστικά με τις αριθμητικές μεταβλητές ή χαρακτηριστικά (διαβάστε καλά τις εκφωνήσεις κάθε ερωτήματος). Όπως θα διαπιστώσετε, θα πρέπει η καθεμιά ή καθένας από εσάς να έχει 2 ή 3 τέτοιες μεταβλητές.

- Classification:
    - **Τυχαία Δέντρα:**
        - Αρχικά χωρίστε το σύνολο δεδομένων σε εκπαιδευτικά και δοκιμαστικά σύνολα. Συνιστάται διαχωρισμός 75% εκπαιδευτικά και 25% δοκιμαστικά.
        - Χρησιμοποιήστε τη συνάρτηση rpart (από την ομώνυμη βιβλιοθήκη) για να δημιουργήσετε ένα μοντέλο δέντρου απόφασης που να περιλαμβάνει όλες τις ανεξάρτητες μεταβλητές ή χαρακτηριστικά (δηλαδή όλα τα χαρακτηριστικά πλην φυσικά της μεταβλητής στόχου class).
        - Οπτικοποιήστε το δέντρο απόφασης και ερμηνεύστε τα αποτελέσματα.
        - Επαναλάβετε τα παραπάνω άλλες 2 φόρες, για δύο διαφορετικές παραμετροποιήσεις του δέντρου απόφασης, όπου επιλέγετε διαφορετικούς συνδυασμούς χαρακτηριστικών για το μοντέλο σας. Περιγράψτε με συντομία τι παρατηρείτε όταν συγκρίνετε τις μετρικές απόδοσης ανάμεσα σε αυτά τα 2 μοντέλα και το αρχικό του προηγούμενου ερωτήματος (που περιελάβανε όλα τα χαρακτηριστικά).
    - **k-Nearest Neighbors (k-NN):**
        - Εφαρμόστε τον αλγόριθμο k-NN χρησιμοποιώντας τη βιβλιοθήκη ‘ class ‘ .
        - Εφαρμόστε k-fold cross-validation χρησιμοποιώντας το πακέτο caret, όπως είδαμε στο εργαστήριο. Χρησιμοποιήστε το trainControl() για να ορίσετε τη μέθοδο επανάληψης. Επιλέξτε τιμή k = 5.
        - Με βάση το προηγούμενο ερώτημα, ποια τιμή k των πλησιέστερων γειτόνων προσφέρει την καλύτερη ακρίβεια; Δικαιολογήστε την απάντησή σας με βάση τα ευρήματα.
    - **Απλό Μοντέλο NN:**

- Χωρίστε και πάλι το σύνολο δεδομένων σε εκπαιδευτικά και δοκιμαστικά σύνολα. Συνιστάται διαχωρισμός 75% εκπαιδευτικά και 25% δοκιμαστικά.
- Χρησιμοποιώντας τη συνάρτηση nnet από την ομώνυμη βιβλιοθήκη (δείτε κώδικα από Εργαστήριο 5, χρησιμοποιήστε τα ίδια arguments *size = 1, linout = FALSE, trace = FALSE* πέρα από το μοντέλο και τα δεδομένα) εκπαιδεύστε και στη συνέχεια ελέγξτε την απόδοση του μοντέλου σας, το οποίο θα πρέπει να περιλαμβάνει όλα τα διαθέσιμα χαρακτηριστικά (να είναι δηλαδή ΄class ~ . ‘).
- Επαναλάβετε το παραπάνω, αυτή τη φορά για διαφορετική μοντελοποίηση (πχ ‘ class ~ Feature_2 + Feature_1 + Feature_6 ΄)

Clustering:

- **k-Means:**
  - Εφαρμόστε τον k-means **στα αριθμητικά χαρακτηριστικά του συνόλου δεδομένων (τα οποία θα πρέπει να κανονικοποιήσετε με τη συνάρτηση scale())** σας χρησιμοποιώντας τη συνάρτηση kmeans().
    - Βοηθητικός κώδικας για την κανονικοποίηση:

    data_numeric = data[1:1000,c(2,7,8)]
    data_numeric <- as.data.frame(scale(data_numeric))

  - Προσδιορίστε τον βέλτιστο αριθμό ομάδων/συστάδων χρησιμοποιώντας τη μέθοδο Elbow ή τη μέθοδο Silhouette (δείτε Εργαστήριο 8).
- **Hierarchical Clustering:**
  - Για υπολογιστικούς λόγους, επιλέξτε τυχαία 1000 δείγματα από το σύνολο δεδομένων σας.
  - Εφαρμόστε Hierarchical clustering **σε 2 αριθμητικά χαρακτηριστικά** της επιλογής σας **(τα οποία θα πρέπει να κανονικοποιήσετε)** χρησιμοποιώντας τη συνάρτηση hclust() (δείτε Εργαστήριο 10 και τα βήματα 1-6 εκεί). Χρησιμοποιήστε για μέθοδο απόστασης την Μανχάταν (hint: distance_matrix <- dist(data, method = "manhattan")).

- Στην μέθοδο όρισμα της hclust ορίστε ‘complete.
- Στο αντίστοιχο βήμα 4 από το Εργαστήριο 10, ‘κόψτε’ το δέντρο στην τιμή k = 2
- Οπτικοποιήστε το παραγόμενο δενδρόγραμμα (hint: δείτε Εργαστήριο 10) και στη συνέχεια οπτικοποιήστε τις συστάδες με χρήση της ggplot (ξανά, δείτε το Εργαστήριο 10 και το βήμα 6).

- **DBSCAN:**
  - Υλοποιήστε το DBSCAN **στα ίδια αριθμητικά χαρακτηριστικά με το παραπάνω** χρησιμοποιώντας το πακέτο dbscan (δείτε Εργαστήριο 8_0).
  - Με παραμέτρους ‘ eps = 0.5, minPts = 2 ‘, πόσες ομάδες/clusters παρήγαγε ο αλγόριθμος; Οπτικοποιήστε τις ομάδες που σχηματίστηκαν (δείτε το Εργαστήριο 8_0). Δοκιμάστε 3 ακόμα διαφορετικές παραμετροποιήσεις.

Μέτρα αξιολόγησης για τα παραπάνω

- Για τα ταξινομητικά μοντέλα, παραθέστε (με πίνακα ή απλό screenshot) τον confusion matrix, την ακρίβεια (accuracy), και την την ευαισθησία (sensitivity) του μοντέλου σας.
- Σχολιάστε τα ευρήματά σας

Έξτρα μπόνους (μέχρι +1 βαθμός) και πολύ δύσκολα ερωτήματα για όσες και όσους παρακολουθούσαν τα εργαστήρια:

Bonus ερώτημα 1.

1. Επιστρέφουμε στο ερώτημα με τα δέντρα απόφασης και την ταξινόμηση. Καλείστε να τροποποιήσετε τη λύση σας και να υλοποιήσετε έναν κάπως διαφορετικό αλγόριθμο. Συγκεκριμένα, θα πρέπει να υλοποιήσετε τον αλγόριθμο Τυχαίου Δάσους (Random Forest https://en.wikipedia.org/wiki/Random_forest) με bagging

https://en.wikipedia.org/wiki/Bootstrap_aggregating, **ΧΩΡΙΣ** την χρήση διαφορετικής ή τρίτης συνάρτησης. Δηλαδή, θα πρέπει να βασιστείτε στη συνάρτηση rpart που χρησιμοποιήσατε παραπάνω και σε βασικές εντολές της R για δειγματοληψία. Ο αριθμός των δέντρων που θα χρησιμοποιήσετε να είναι 1000.

Bonus ερώτημα 2.

1. Μελετήστε καλά τα αποτελέσματά σας για την ταξινόμηση στην εργασία σας (Δέντρα απόφασης, kNN, νευρωνικό δίκτυο και όλες οι διαφορετικές εκδοχές/μοντέλα που υλοποιήσατε). Ποιο είναι το τελικό συμπέρασμα στο οποίο οδηγήστε ως προς τη σημαντικότητα των χαρακτηριστικών; Ποια χαρακτηριστικά 'βοήθησαν' περισσότερο τους αλγόριθμούς σας; Αναπτύξτε αναλυτικά το σκεπτικό σας και δικαιολογήστε την απάντησή σας.
2. Κάντε το ίδιο σχετικά α) με το ποια χαρακτηριστικά δεν συνεισέφεραν αρκετά και β) ποια χαρακτηριστικά παρουσιάζουν εξάρτηση μεταξύ τους.

2ο μέρος (10% του βαθμού της εργασίας)

- Συζητήστε σε 1-2 παραγράφους τις διαφορές στις προβλέψεις α) από τα ταξινομητικά/classification μοντέλα και β) από την ομαδοποίηση/clustering.
- Προτείνετε πιθανές βελτιώσεις ή εναλλακτικές τεχνικές που θα μπορούσαν να εφαρμοστούν στο σύνολο δεδομένων.